1 A Biologically Inspired Model for the Simultaneous Recognition of Identity and Expression

DONALD NETH & ALEIX M. MARTINEZ
Dept. Electrical and Computer Engineering
Dept. Biomedical Engineering
The Ohio State University

1.1 INTRODUCTION

Faces provide a wide range of information about a person's identity, race, sex, age and emotional state. In most cases, humans easily derive such information by processes that appear rapid and automatic. However, upon closer inspection one finds these processes to be diverse and complex. In this chapter, the perception of identity and emotion is examined. We argue that the two cannot be studied independently of each other because the internal computational processes are intertwined. Next, a computational model is developed for the processing of expression variant face images. This model is then applied to matching the identity of face images with differing expression. Finally, the model is used to classify expressions from face images.

1.1.1 Perception of Identity and Emotion

Historically, research on human face perception has taken two perspectives. The first involves the recognition of individual identity through perception of the face. The second involves the perception of emotional expression from the face. This division between facial recognition and perception of emotional expression has its roots in clinical observations. In a condition known as *prosopagnosia*, patients are unable to recognize faces of people familiar to them while perception of expression remains relatively intact [1]. This condition typically results from brain injury or stroke, however, some cases of congenital prosopagnosia have been observed [5]. The preservation of emotional perception amidst a loss of recognition suggests that two relatively independent systems are involved in face perception. A condition known as Capgra's delusion offers complementary support to this theory [23]. Capgra's delusion is the belief that significant others are no longer who they were. Instead, they are thought to have been replaced by doubles, impostors, robots, or aliens. Patients with Capgra's delusion are able to recognize faces; however, they deny their authenticity. While prosopagnosia is the inability to recognize previously familiar faces, there is some evidence that the ability to discriminate familiar faces is retained outside of conscious awareness – a normal elevated skin conductance is observed in response to familiar faces. Thus, prosopagnosia can be characterized as a failure in conscious face recognition coupled with an intact unconscious or covert mechanism. Conversely, Capgra's delusion may be characterized as an intact conscious face recognition system coupled with a failure in covert or unconscious recognition. The anomalous perceptual experiences arising from failure of the covert processing system must be explained by the individual to him or herself – the delusions arise as an attempt to explain or make sense of an abnormal experience.

While prosopagnosia and Capgra's delusion offer compelling illustrations of two major facets of face perception, there is still considerable debate as to the level of independence between face recognition and perception of expression. Nonetheless, past research has tended to treat these two aspects of face perception as involving relatively separate systems. Hence, prior research in face recognition and perception of emotional expression will first be reviewed separately. Finally, an attempt will be made to reconcile the two perspectives.

1.2 FACE RECOGNITION

There is evidence suggesting that the ability to accurately recognize faces relies on an innate cortical face module. The existence of discrete regions of the cerebral cortex specialized for face perception was investigated by Kanwisher, McDermott, and Chun [41]. Face perception was defined to include any higher-level visual processing of faces ranging from the detection of a face as a face to the extraction of information relative to identity, gaze, mood, or sex. A region of the fusiform gyrus that responded differentially to passive face viewing compared to passive object viewing was found and has been subsequently described as the *fusiform face area* (FFA). This region did not simply respond to animal or human images or body parts but to faces in particular. Additionally, this region generalizes to respond to images of faces taken from a different viewpoint with considerably different low-level features from the original set of face images.

In contrast to the notion that humans are hardwired for face perception, it has been suggested that face recognition is a natural consequence of extensive experience with faces. As a model of this process, a *flexible process map* was proposed by Tarr and Gauthier [64]. This model is based on the observation that a number of extrastriate areas are involved in visual object processing and recognition. Through experience with particular visual geometries, associations arise linking task-appropriate recognition strategies that automatically recruit components of the process map. The strong activation of the fusiform gyrus in face processing is thought to be a result of the extensive experience with faces common to all humans. Recognition of faces typically occurs at an individual or sub-category level. In contrast, most objects encountered in daily life are differentiated at a category level. It is suggested that the fusiform gyrus represents a cortical area in which sub-category discrimination occurs. The ubiquitous experience with faces, and the subsequent development of expertise, is reflected in the activation of the fusiform gyrus in face processing. However, this should not be interpreted to mean the fusiform gyrus is a dedicated module for face processing alone. In a study of the involvement of the fusiform gyrus with expert-level discrimination, subjects were trained to discriminate novel objects known as greebles [29]. Functional imaging revealed increased activation of the right fusiform gyrus during expert-level discrimination tasks. The authors hold that expertise is an important factor leading to the specialization of the fusiform gyrus in face processing.

An interactive specialization view in which cortical specialization is an emergent product of the interaction of both intrinsic and extrinsic factors has been proposed [18]. According to this view, the development of face recognition relies on two processes. The first process, termed *Conspec*, is a system operating from birth that biases the newborn to orient toward faces. It is mediated by primitive subcortical circuits. The second process, termed *Conlern*, relies on a system sensitive to the effects of experience through passive exposure to faces. It is mediated by the developing cortical circuits in the ventral visual pathway. Newborn infants show evidence of recognizing facial identity. Before the specialization of cortical circuits, face stimuli are processed in the same manner as other visual stimuli. This ability is then augmented as the Conlern system emerges at around 6 to 8 weeks of age. Newborns also exhibit preferential tracking, however, it is not specific to the fine details of facial features but relies on the arrangements of the elements comprising the face, i.e. configural information. This preferential tracking declines sharply at 4 to 6 weeks of age, similar to the decline seen in other reflex-like behaviors thought to be due to inhibition by developing cortical circuits. In addition, it has been observed that newborns orient to patterns with a higher density of elements in the upper visual field. Infants also demonstrate the ability to recognize individual facial identity, gazing longer at the mother's face than the face of a stranger. Cortical development impacts the mental representation of facial identity. This representation is thought to be multi-dimensional in that it encodes multiple aspects of a face. While the specific dimensions of this face space are not clearly delineated, they are thought to be learned rather than being pre-specified at birth. For an infant, the face space will contain fewer entries during development than in adulthood. Furthermore, infants and children are less likely to use a large number of dimensions since relatively few are required to distinguish the smaller number of faces in their environments. Infants develop a face-space at around 3 months of age and begin to form categories based on the faces they see. The authors offer an alternative to the view that face processing is merely an example of acquired expertise. They propose that it is special in that the timing of particular visual inputs during development is critical for normal development. The regions of the ventral occipito-temporal cortex have the potential to become specialized for face recognition but require experience with faces for the specialization to arise.

A similar conclusion is drawn by Nelson [54] who characterizes face recognition as an important adaptive function that has been conserved across species. Monkeys have been observed to utilize a process similar to that of human adults when studying faces – the internal parts of the face are more significant than the external parts. In humans, as in monkeys, it is adaptive for the young infant to recognize potential caretakers. At around 4 months of age, human infants exhibit superior recognition performance for upright faces versus inverted faces. This suggests that they have developed a schema for faces and have begun to view faces as a special class of stimuli. Between 3 and 7 months, the ability to distinguish mother from stranger becomes more robust. It is held that the development of face recognition is an *experience-expectant* process. Such a process refers to the development of skills and abilities that are common to all members of the species and depends on exposure to certain experiences occurring over a particular period of time in development. The involvement of the inferotemporal cortex in face recognition may have been selected for through evolutionary pressures or the properties of the neurons and synapses in this region may be particularly tuned to the task of face recognition. Such specialization occurs rapidly within the first months of life. As experience with faces increases, perceptual learning leads to further specialization of this area.

1.2.1 Configural Processing

Maurer, Le Grand, and Mondloch [52] identified three types of configural face processing: (1) sensitivity to first-order relations, (2) holistic processing, and (3) sensitivity to second-order relations. Sensitivity to first-order relations refers to the ability to identify a stimulus as a face based on the basic configuration of two eyes above a nose above a mouth. Holistic processing refers to the largely automatic tendency to process a face as a gestalt rather than isolated features. Second order relations refer to the distance among internal features.

The role of the fusiform face area in the extraction of configural information from faces and non-faces was investigated by Yovel and Kanwisher [72]. Two tasks were matched for overall difficulty: subjects were asked to discriminate sequentially presented image pairs of faces or houses that could differ in (1) only the spatial relation between parts and (2) only in the shapes of the parts. The study was extended by repeating the approach with inverted image pairs. Thus, four test conditions were used. The FFA showed a significantly higher activation to faces than houses but showed no difference between the part and configuration tasks. The inversion effect was absent for houses but equally strong for part and configuration tasks for houses. The authors conclude that face perception mechanisms are domain specific (i.e., engaged by faces regardless of processing type) rather than process specific (i.e., engaged in specific process depending on task type). The similar results for configuration and part tasks is contrary to the commonly held view that face processing is largely configural and that the inversion effect results from the disruption of configural processing.

1.2.2 Cognitive Model of Face Recognition

In an influential paper, Bruce and Young [11] proposed a functional model to describe the process of face recognition, Figure 1.1. An abstract visual representation must be established to mediate recognition even though an identical image is rarely viewed on successive occasions. This demonstrates an ability to derive structural codes that capture aspects of the facial structure essential to distinguish one face from another. Visually derived semantic codes are useful in describing such factors as age and sex. Identity-specific semantic codes contain information about a person's occupation, where he might be encountered, etc. It is the recovery of identity-specific semantic codes that creates the feeling of knowing. A name code exists independently of the identity-specific code. Face recognition units (FRU) are proposed that contain stored structural codes describing each face known to a person. FRUs can access identity-specific semantic codes held in associative memory known as person identity nodes. Names are accessed only through the person identity nodes. Names are thought to be abstract phonological representations stored separately from semantic representations but accessed through them.

Ellis [22] employed the Bruce and Young model to explain various blocks in facial recognition. A temporary block between semantic and name retrieval is responsible for errors in which the perceiver can remember everything about a person except his/her name. This theory disallows errors in which the person's name is known but the semantic information is not. While parallel access to semantics and spoken word-forms is believed to occur for the recognition of familiar written words, this does not seem to be the case for faces. Several factors influence the speed and accuracy of familiar face recognition. In repetition priming, a face will be recognized as familiar more quickly when it has been viewed previously. Repetition priming has been observed primarily in familiarity decisions and is not found in expression or sex decisions. Another factor affecting face processing is the distinctiveness of the face. In familiarity tasks, distinctiveness results in faster recognition than with more typical faces. Finally, in associative priming, past co-occurrences between familiar

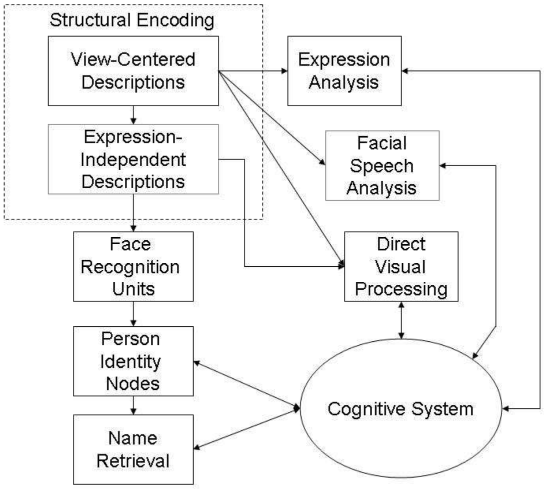


Fig. 1.1 Cognitive model of face processing (Adapted from Bruce & Young, 1986).

stimuli allow one to predict the recognition of the other. For example, if a face to be recognized is preceded by a related and familiar face, recognition will occur more rapidly.

1.3 FACIAL EXPRESSION OF EMOTION

Ekman and colleagues [21], expanding on the work of Darwin [17], identified six universal emotional expressions: anger, sadness, fear, surprise, happiness, and disgust. Ekman [19] extended his earlier work by compiling a list of basic emotions including amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame. According to Ekman, these emotions share basic characteristics. They are distinctive universal signals with specific physiological concomitants. They provide automatic appraisal, in particular, to distinctive universals in antecedent events. Emotions typically have a quick onset which is essential for their adaptive value. It is also adaptive for the response changes to be of brief duration unless the emotion is evoked again. In an effort to quantify facial expressions, Ekman and Friesen [21] developed the Facial Action Coding System (FACS). Facial expressions were extensively examined and their component motions were determined. Forty-four separate action units (AUs) were identified with five levels of intensity ranging from A to E. Thirty AUs are related to contraction of specific facial muscles. Eight AUs describe different movements related to head orientation, while four AUs are used to describe eye direction. While the use of FACS allows a high degree of refinement in classifying facial expressions, Ekman's approach to emotion and its expression remains categorical [19].

In contrast, a number of researchers have proposed continuous models of emotion and facial expressions [68, 59, 60, 58, 13, 61]. The circumplex model proposed by Russell [59] represents emotions in two dimensions reflecting pleasure and arousal. The horizontal dimension ranges from extreme displeasure (e.g., agony) to extreme pleasure (e.g., ecstasy). The vertical dimension ranges from sleep at one extreme to frenetic excitement at the other. These two dimensions form a continuous space without clearly delineated boundaries or prototypical emotions. This model plays a central role in Russell's [61] concept of *core affect* which is the neurophysiological state accessible as the simplest, non-reflective feelings. Core affect is always present but may subside into the background of consciousness. The prototypical emotions described in categorical approaches map into the core affect model but hold no special status. According to Russell [61], prototypical emotions are rare; what is typically categorized as prototypical may in fact reflect different patterns than other states classified as the same prototype. He states that emotional life comprises the continuous fluctuations in core affect, the on-going perception of affective qualities, frequent attribution of core affect to an external stimulus, and instrumental behaviors in response to that external stimulus. The many degrees and variations of these components will rarely fit the pattern associated with a prototype and will more likely reflect a combination of various prototypes to varying extents. While Russell makes a compelling argument for the continuous nature of emotion, humans appear predisposed to experience certain continuously varying stimuli as belonging to distinct categories.

Facial expressions are controlled by both pyramidal and extrapyramidal tracts which provide voluntary and automatic control, respectively. Voluntary control over facial muscles is considered a hallmark of human nonverbal expression and may be due to the articulatory demands of human language [62]. However, there are notable differences between posed and spontaneous expressions. Such differences are particularly evident in smiling. The Duchenne smile is described as the combined contraction of the *zygomaticus major* and *orbicularis oculi* muscles and is thought to occur with spontaneously occurring enjoyment [20]. False smiles are described as those made to convince another that enjoyment is occurring when it is not. Masking smiles are made to conceal negative emotions. Miserable smiles denote a willingness to endure an unpleasant situation. The Duchenne smile was found to occur during solitary enjoyment and was associated with greater left-hemisphere anterior temporal and parietal activation compared to other smiles [20]. Differences in the dynamic features of social and spontaneous smiles were investigated by Cohn and Schmidt [16]. Spontaneous smiles exhibit characteristics of automatic movement. Automatic movements are thought to be pre-programmed and are characterized by a consistent relationship between maximum duration and amplitude of movement. Posed (social) smiles exhibit a far less consistent relationship between duration and amplitude. Smiles comprise an initial onset phase, a peak, and an offset phase. The onset phase was used in this study because it provides the most conspicuous change in the face as perceived by human observers. Amplitude was found to be smaller in spontaneous smiles than in social smiles. Timing and amplitude measures were used in a linear discriminant classifier resulting in a 93% recognition rate. With timing measures alone, the recognition rate was 89%.

Gallese, Keysers, and Rizzolatti [28] suggest that mirror mechanisms in the brain allow the direct understanding of the meaning of action and emotions of others by internally replicating (or simulating) them without any reflective mediation. Thus, conceptual reasoning is not necessary for such understanding. When action is observed, there is concurrent activation of part of the same motor areas used to perform the action. Similarly, it is thought that mirror mechanisms allow individuals to simulate the emotional state of others. Within the cerebral cortex, the superior temporal sulcus (STS) is activated by observation of movements of the eyes and head, movements of the mouth, and meaningful hand movements. Some groups of cells respond preferentially to hand movements. Typically, the groups respond better to a particular kind of hand movement. The responsiveness of the cell group is independent of the object acted upon and the speed at which the hand moves. Additionally, the responsiveness of the cell group is greater when the movement is goal-directed. Observation of whole body movements activates a posterior region of the STS. The STS is also activated by static images of the face and body. Taken together, this suggests that the STS is sensitive to stimuli that signal the actions of another individual [2].

Valentine [66] makes the distinction between identification and recognition in that identification requires a judgment pertaining to a specific stimulus while recognition requires only a judgment that the face has been seen before. He posits the capability to reliably distinguish friend from foe would confer an evolutionary advantage over simply knowing that a face has been seen before.

1.3.1 Categorical Perception of Expression and Emotion

Categorical perception is a psychophysical phenomenon which may occur when a set of stimuli ranging along a physical continuum is divided into categories. Categorical perception involves a greater sensitivity to changes in a stimulus across category boundaries than when the same change occurs within a single category [31]. Categorical perception has been observed in a variety of stimuli including colors [10], musical tones [43], among many others. There is significant evidence that facial expressions are perceived as belonging to distinct categories. In a study by Calder *et al.* [14], the categorical perception of facial expressions based on morphed photographic images was investigated. Three expression continua were employed: happiness-sadness, sadness-anger, and anger-fear. Subjects were first asked to identify the individual stimuli by placing them along particular expression continua. Subjects were then asked to perform a discrimination task in which stimuli A, B, and X were presented sequentially. Subjects were asked whether X was the same as A or B. Results indicate that each expression continuum was perceived as two distinct categories separated by a boundary. It was further found that discrimination was more accurate for across-boundary rather than within-boundary pairs.

In a classic study, Young *et al.* [71] investigated whether facial expressions are perceived as continuously varying along underlying dimensions or as belonging to discrete categories. Dimensional approaches were used to predict the consequences of morphing one facial expression to another. Transitions between facial expressions vary in their effects, depending on how each expression is positioned in the emotion space. Some transitions between two expressions may involve indeterminate regions or a third emotion. In contrast, a transition from one category to another may not involve passing through a region which itself may be another category. In this case, changes in perception should be abrupt. Four experiments were conducted using facial expressions from the Ekman and Friesen series [21]. All possible pairwise combinations of emotions were morphed and presented randomly. Subjects identified intermediate morphs as belonging to distinct expression categories corresponding to the prototype end-points. No indeterminate regions or identification of a third emotion were observed. This supports the view that expressions are perceived categorically rather than by locating them along underlying dimensions. The authors suggest that categorical perception reflects the underlying organization of human categorization abilities.

1.3.2 Human Face Perception – Integration

Human face perception engages the visual system in processing multiple aspects of faces including form, motion, color, and depth. Visual information processing in humans has predominantly developed along two fairly segregated pathways: one for form and another for motion. Projections from the primary visual cortex form a *dorsal stream* that progresses to portions of the middle temporal lobe (MT), the medial superior temporal area (MST), and portions of the parietal lobe associated with visuospatial processing. The *ventral stream*, which is associated with object recognition tasks involving texture and shape discrimination, comprises projections from the primary visual cortex to the inferior temporal cortex. Behavioral and lesion studies support a functional distinction between the ventral and dorsal streams with motion processing occurring primarily in the dorsal stream and shape discrimination occurring primarily in the ventral stream. Ungerleider, Courtney, and Haxby [65] suggest that the functional distinction extends to the prefrontal cortex and the working memory system: ventrolateral areas are involved primarily in working memory for objects while the dorsolateral areas are primarily involved with spatial working memory. The organization of the human visual systems reflects the importance of form and motion in the processing of visual information. Thus, it is reasonable to consider face processing from this perspective.

Many neural structures are involved in both recognition and perception of facial expressions. A distributed neural system for face perception, including bilateral regions in the lateral inferior occipital gyri (IOG), the lateral fusiform gyrus (LFG), and posterior superior temporal sulcus (STS) was investigated in an fMRI study by Hoffman and Haxby [35]. It was found that the representation of face identity is more dependent on the IOG and LFG than the STS. The STS is involved in perception of changeable aspects of the face such as eye gaze. Perception of eye gaze also activated the spatial recognition system in the intraparietal sulcus which was thought to encode the direction of the eye gaze and to focus attention in that direction. These findings were integrated into a distributed neural model for face perception formulated by Haxby, Hoffman and Gobbini [32]. The core system comprises three brain regions involved in separate but inter-connected tasks, Figure 1.2. The lateral fusiform gyrus processes invariant aspects of faces and is involved in the perception of identity. The STS processes the dynamic aspects of faces including expression, eye gaze, and lip movement. The inferior occipital gyri are involved in the early perception of facial features. A key concept of the model is that face perception is accomplished through a coordinated participation of multiple regions.

According to the Bruce and Young model, recognition and expression processing function independently. Much of the research reviewed above suggests a partial independence of the two processes. However, complete independence is unlikely [15, 56] and it remains to be determined whether their interaction is direct or indirect [49]. Viewed from the perspective of the visual system, invariant aspects of the face should engage the ventral stream while variable aspects or motions should engage the dorsal stream. There is also evidence that invariant aspects of the face facilitate recognition of a person's identity while the variable aspects allow inferences regarding that person's state of mind. However, Roark *et al.* [57] suggest that supplemental information can be derived from facial motion in the form of dynamic facial signatures which can augment recognition of familiar faces. Such signatures should be processed by the dorsal visual stream ultimately engaging the STS. The authors also speculate that motion can be useful in enhancing the representation of invariant aspects of the face. Such structure-from-motion analysis also engages the dorsal stream. They suggest that these concepts be integrated into Haxby's distributed neural model.

1.4 MODEL OF EXPRESSION-VARIANT PROCESSING

While Haxby's distributed neural system is a fairly comprehensive model, it is still not clear by which mechanism the brain successfully accomplishes the

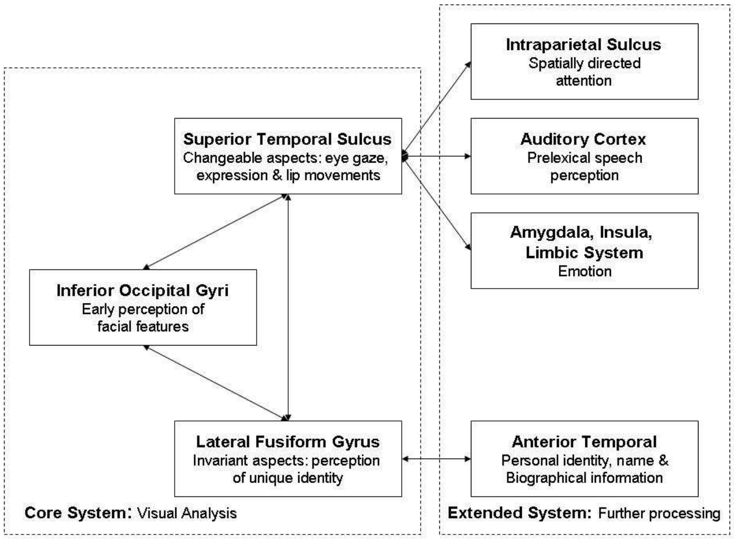

Intraparietal Sulcus
Spatially directed attention
Superior Temporal Sulcus
Changeable aspects: eye gaze, expression & lip movements
Auditory Cortex
Prelexical speech perception
Amygdala, Insula, Limbic System
Emotion
Inferior Occipital Gyri
Early perception of facial features
Lateral Fusiform Gyrus
Invariant aspects: perception of unique identity
Anterior Temporal
Personal identity, name & Biographical information
Core System: Visual Analysis
Extended System: Further processing

Fig. 1.2 Distributed neural system for face perception. (Adapted from Haxby, Hoffman, & Gobbini, 2000).

matching of two or more face images when differences in facial expression make the (local and global) appearance of these images different from one another. There seems to be a consensus that faces are processed holistically rather than locally, but there is not yet consensus on whether information on facial expression is passed to the identification process to aid recognition of individuals or not. As mentioned in the previous section in this chapter, some models proposed in the past suggest that to recognize people's identity we use a process that is completely decoupled from that of recognizing the facial expression [11]. Others propose that a connection must exist between the two processes [30]. Psychophysical data exist in favor of and against each view [12, 44, 24, 25, 70, 63, 4]. Martinez [49] posed a fundamental question in face recognition: *Does the identification process receive information from or interact with the process of facial expression recognition to aid in the recognition of individuals?* It has been noted that subjects are slower in identifying happy and angry faces than faces with neutral expressions. A model was proposed in which a motion estimation process is coupled with the processing of the invariant aspects of the face. Both of these processes contribute to recognition of identity and the perception of facial expressions. The key element is the addition of a deformation of the face (*DF*) module which calculates the apparent physical deformation between faces by computing the motion field between faces to be matched. A separate module processes the invariant aspects of the face. Both processes occur in tandem and the outputs of both are

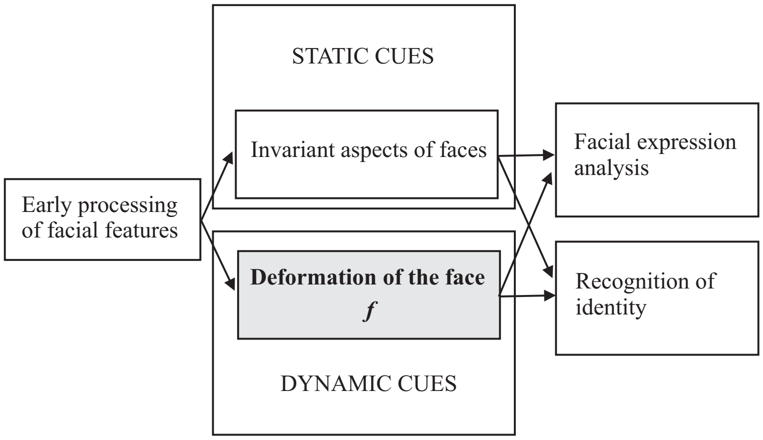

STATIC CUES
Invariant aspects of faces
Early processing of facial features
Deformation of the face
f
DYNAMIC CUES
Facial expression analysis
Recognition of identity

Fig. 1.3 Depiction of the different processes of the model presented in this chapter. The modules dedicated to the recognition of identity and expression are dissociated, although they both obtain information from the common process *DF* (Deformation of the Face) and from the processes that compute static cues. This is key to explain the psychophysical date described in the past.

fed into two independent processes: one for the recognition of identity and the other for analysis of expression. The addition of the DF module offers an explanation for increases in recognition time for faces with larger deformations - the more complex the deformation, the longer it takes to process the implied motion. According to this model, it is now logical to expect the identification of faces to be slower for those cases where a larger deformation of the face exists, since we need to go through the motion estimation module *DF*. The more complex the facial expression, the more time (generally) needed to compute an approximation of the muscle activity of the face (see *Results*). For example, the *DF* module shown in Fig. 1.3 does not need to estimate any motion for the neutral facial expression case, but requires the computation of an approximation of the motion of a smile and other facial expressions.

We hypothesize that the motion field is necessary (or, at least, useful) to successfully match the local and global features of a set of faces when those bear distinct facial expressions. It has been shown that motion plays an important role in recognizing identity and facial expressions in a sequence of images [34, 67, 45]. In addition, uncommon deformations or uncommon sampling times disrupt identification of individuals [33] and of facial expressions [40]. This seems reasonable, because the local texture of a face changes considerably as the facial expression also changes. Facial expressions change the local texture of each of local face areas which appear quite distinct under different expressions. A classification (or identification) algorithm, however, will need to find where the invariant features are. The motion field (deformation) between the two images that one wants to match can be used to determine the features that have changed the least between the two images, and thus, which are the best candidates for matching purposes.

1.4.1 Recognition of Identity

Consider two different images, $\mathbf{I}_1$ and $\mathbf{I}_2$, both of n pixels. We can redefine the images as vectors taking values in an n-dimensional space. We shall denote this as $\mathbf{V}_1$ and $\mathbf{V}_2$, with $\mathbf{V}_i \in \mathbb{R}^n$. The advantage of doing this is that it allows comparisons of the images by means of vector operations such as subtraction

$$\|\mathbf{V}_1 - \mathbf{V}_2\| \tag{1.1}$$

where $\| \cdot \|$ denotes the L_2 norm (i.e., Euclidean distance). In this definition stated here, we assume that all faces have been aligned (with respect to the main facial features) in such a way that the eyes, mouths, noses, etc. of each of the images are at roughly the same pixel coordinates, e.g. [6, 48]. The approach defined above, in Eq. (1.1), has proven to perform well when frontal face images with similar facial expressions are compared to each other. However, this comparison becomes unstable when matching face images bearing different facial expressions [48]; hence pixels can now carry information of different features.

The incorporation of the *DF* process in our model allows us to represent face processing as

$$\|f^{-1}(\mathbf{V}_1 - \mathbf{V}_2)\|, \tag{1.2}$$

where f is a function proportional to the motion of each pixel, i.e., the movement representing the facial expression of the test image. Intuitively, f is a function that keeps correspondences between the pixels of the first and second images. Eq. (1.2) can be interpreted as follows; pixels (or local areas) that have been deformed largely due to local musculature activity will have a low weight, whereas pixels that are less affected by those changes will gain importance. We can formally define f^{-1} as taking values linearly inverse to those of f, i.e.,

$$MAX_F - \|\mathbf{F}_i\| \tag{1.3}$$

where $\mathbf{F}$ is the motion flow (i.e., motion between two images), $\mathbf{F}_i$ the motion vector at the ith pixel, and $MAX_F = max_{\forall i}\|\mathbf{F}_i\|$ (the magnitude of the largest motion vector in the image).

The value of f corresponds thus to the outcome of the *DF* process. Note that f defines the face deformation (motion) between two images and, therefore, can also be used to estimate the facial expression of a new incoming face image. As mentioned earlier, experimental data supports this belief.

1.4.2 Motion estimation

Visual motion between two images can be expressed mathematically by local deformations that occur in small intervals of time, δt, as

$$\mathbf{I}(x,y,t) = \mathbf{I}(x + u\delta t, y + v\delta t, t + \delta t), \tag{1.4}$$

where $\mathbf{I}(x,y,t)$ is the image value at point (x,y) at time t, (u,v) are the horizontal and the vertical image velocities at (x,y) and δt is considered to be small [36]. We note that in our model $f = (u,v)$.

If we assume that the motion field (i.e., the pixel correspondences between the two images) is small at each pixel location, the motion estimator can be represented by the first-order Taylor series expansion as

$$E_D = \int\int \rho\left(\mathbf{I}_x u + \mathbf{I}_y v + \mathbf{I}_t\right) dxdy \tag{1.5}$$

where $(\mathbf{I}_x, \mathbf{I}_y)$ and $\mathbf{I}_t$ are the spatial and time derivatives of the image, and ρ is an *estimator*.

To resolve the above equation, it is necessary to add an additional constraint. The most common one is the spatial coherence constraint [36], which embodies the assumption that neighboring pixels in an image are likely to belong to the same surface and, therefore, a smoothness in the flow is expected. The first-order model of this second constraint is given by

$$E_S = \int\int \rho\left(\nabla(u,v)\right) dxdy \tag{1.6}$$

where ∇ represents the gradient.

Visual motion is determined by minimizing the regularization problem

$$E = E_D + \lambda E_S. \tag{1.7}$$

Although the objective function E is non-linear (and a direct solution does not exist for minimizing it), a convex approximation can be obtained [9]. The global minimum can then be determined iteratively.

This procedure is most effective when the object displacements between consecutive images are small. When object displacements are large, a coarse-to-fine strategy can be used. In the current work, the pyramid method of [8] was used. In order to satisfy the small-displacement assumption we begin with a reduced-resolution representation of the images. The optical flow is computed for the low-resolution images and then projected to the next level of the pyramid where the images in the sequence have a higher resolution. At each level of the pyramid, the optical flow computed from the previous level is used to warp the images in the sequence. This process is repeated until the flow has been computed at the original resolution. The final flow field is obtained by combining the flow information of each of the levels of the

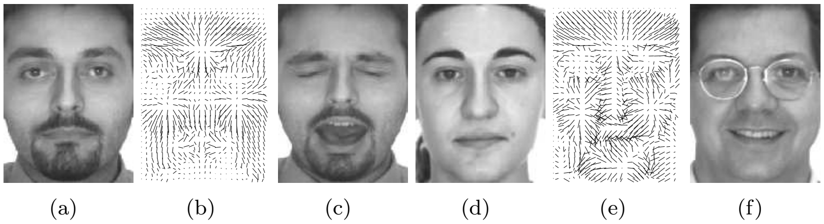


Fig. 1.4 The motion estimation in (b) defines the muscle movement defining the expression change between (a) and (c). The motion estimation in (e) defines the expression and identity change between (d) and (f).

pyramid. The number of levels on the pyramid will be dictated by the largest motion in the sequence of images.

The approach to motion estimation defined above may result in biased results whenever the scene's illumination in the sample and test face images are distinct. To resolve this issue, one can include the modeling of this illumination variation in Eqs. (1.4-1.5). Negahdaripour [53] extends the above definition of motion flow to include radiometric changes into its computation. This definition requires a extension of the 2-D motion field vector (u, v) to a 3D transformation field given by $(u, v, \delta e)$. The last component, δe, describes the radiometric transformation of the image sequence. This provides us with a new model for the motion, given by

$$\mathbf{I}(x + u\delta t, y + v\delta t, t + \delta t) = \mathbf{M}(x, y, t)\mathbf{I}(x, y, t) + \mathbf{C}(x, y, t), \tag{1.8}$$

in which the brightness at a pixel in two consecutive images is related via the motion parameters u and v and the radiometric parameters $\mathbf{M}$ and $\mathbf{C}$, as shown. Here, $\mathbf{M}$ defines light changes resulting in multiplier variations, e.g., change in homogenous or non-homogeneous intensity, while $\mathbf{C}$ defines additive terms, such as cast shadows.

The data conservation constraint corresponding to (1.8) is

$$E_D = \int\int \rho(\mathbf{I}_x u + \mathbf{I}_y v + \mathbf{I}_t - (\mathbf{I}\mathbf{m}_t + \mathbf{c}_t)), \tag{1.9}$$

where $\mathbf{m}_t$ and $\mathbf{c}_t$ are the time derivatives of $\mathbf{M}$ and $\mathbf{C}$, respectively. Since, we now have two more variables to estimate, the smoothness constraint, needs also to include the following minimizations

$$E_M = \int \rho(\nabla \mathbf{m}_t) \quad \texttt{and} \quad E_C = \int \rho(\nabla \mathbf{c}_t). \tag{1.10}$$

By using a robust function for $\rho(.)$, we can generate very precise estimates of the motion (i.e., *DF*) even under varying illumination, as we have shown in [42]. Two examples of motion estimation between face images are shown in Fig. 1.4.

1.4.3 Recognition of Expression

In our model, each facial expression is classified in a category according to the motion field of the face, f. The direction of motion is used to determine the class [3], while the magnitude of the motion can be used to specify the intensity of a given expression. These two parts of the motion can be expressed mathematically as

$$S_{M_i} = abs(\|\mathbf{F_t}_i\| - \|\mathbf{F_p}_i\|) \quad and \quad S_{A_i} = arccos\frac{\langle \mathbf{F_t}_i, \mathbf{F_p}_i \rangle}{\|\mathbf{F_t}_i\| \|\mathbf{F_p}_i\|}, \qquad (1.11)$$

where $\mathbf{F_t}_i$ and $\mathbf{F_p}_i$ are the vector flows of the two expressions to be compared at the i^{th} pixel, $\langle \mathbf{a}, \mathbf{b} \rangle$ represents the dot product of $\mathbf{a}$ and $\mathbf{b}$, S_{M_i} is the similarity between the magnitude of the i^{th} pixel in the two image flows, and S_{A_i} the similarity between the angles of the two vectors at pixel i.

While this method is normally used to compare two images (i.e. matching), it can also be used to classify (or identify) facial expressions within a group of pre-learned categories. This comparison can be carried out at each pixel location or at specific areas that are known to be most discriminant for a given expression. We can formally express this as

$$S_M = \sum_{i=1}^{m} S_{M_i} \quad and \quad S_A = \sum_{i=1}^{m} \frac{S_{A_i}}{m_o}, \qquad (1.12)$$

where m is the number of pixels where comparison takes place, $m \leq n$, and m_o is the total number of vectors in m with magnitude greater than zero. Note that since the angle similarity can only be computed between actual vectors (of magnitude greater than zero), it is necessary to normalize S_A by the number of comparisons to prevent biases towards images with associated small motions.

In order to appropriately select the value of m, it is convenient to search for those features (i.e. pixels) that best discriminate between categories and those that are most stable within categories. This can be accomplished by means of Fisher's Linear Discriminant Analysis (LDA) [27] and variants [51, 75]. Formally, we define the within and between class scatter matrices of LDA as [27],

$$\mathbf{S}_W = \sum_{j=1}^{c} \sum_{i=1}^{N_j} (\mathbf{v}_{i,j} - \mu_j)(\mathbf{v}_{i,j} - \mu_j)^T \quad and \quad \mathbf{S}_B = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T, \qquad (1.13)$$

where $\mathbf{S}_W$ is the within-class scatter matrix, $\mathbf{S}_B$ is the between-class scatter matrix, c is the number of classes, N_j is the number of samples for class j,

$\mathbf{v}_{i,j}$ is the i^{th} sample of class j, μ_j is the mean vector of class j, and μ is the mean of all classes.

Due to singularity problems, it is generally difficult to compute the LDA transformation of large face images [51]. Additionally, $\mathbf{S}_B$ limits us to a maximum of $c - 1$ dimensions (where c is the number of classes) [75]. Since we usually deal with small values of c and it is known that LDA may perform poorly if the dimensionality of the space is small [47], it is convenient to only use that information which directly specifies the usefulness of each pixel. This is represented in the variances of each feature (pixel) within $\mathbf{S}_W$ and $\mathbf{S}_B$, which is given by the values at the diagonal of each of these matrices:

$$\hat{\mathbf{S}}_W = diag(\mathbf{S}_W) \quad \texttt{and} \quad \hat{\mathbf{S}}_B = diag(\mathbf{S}_B). \tag{1.14}$$

By first finding those pixels (areas) of the face that are most different among classes ($\hat{\mathbf{S}}_B$) and, then selecting those that are most similar across samples of the same class ($\hat{\mathbf{S}}_W$), we can build a classifier that computes the values of S_A in a smaller set of pixels. The result is a classifier that is generally more robust and efficient than one that uses all the pixels of the image.

This model allows us to predict that classification of faces into very distinct categories (e.g. happy and neutral) will be easier than when the two facial expressions are alike (e.g. angry and neutral). As a consequence, response times should be smaller for more distinct classes than for more similar classes. Since the model uses the *DF* procedure described above, we can also predict that when classifying faces within two distinct groups, those that involve larger motions will usually have longer RT. Similarly, those facial expressions that are more difficult to be classified or are more alike, will require the analysis of additional local parts – resulting in longer RT. When a face cannot be reliably classified within one of the categories by looking at the most discriminant areas, we will need to extend our comparison to other areas of the face.

1.5 EXPERIMENTAL RESULTS: RECOGNITION OF IDENTITY

1.5.1 Computational Model

We present an analysis of the computational model defined above. This analysis will facilitate a later comparison with human performance.

1.5.1.1 Performance It is now possible to test two important points advanced in the previous section: *a)* how the suppression of the *DF* process would affect the identification of known individuals, and *b)* how the identification of happy and angry faces is now slower than the recognition of neutral expression faces. In these experiments, we will use the face images of 100

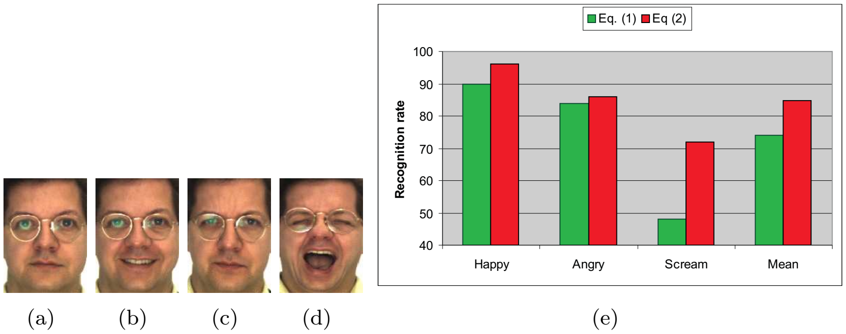


Fig. 1.5 Examples of the following expressions: (a) neutral, (b) happy, (c) angry, and (d) scream. (e) The recognition rates obtained when matching: *i)* happy and neutral faces, *ii)* angry and neutral faces, and *iii)* cream and neutral faces. Note that when we incorporate the information of the *DF* process in our model (i.e. f), the results improve and the matching process becomes less sensitive to the differences in facial expression. Eq. (1) indicates a simple Euclidean distance, and Eq. (2) the weighted measure given in (1.2). Adapted from [50].

individuals of the AR database [46]. The images of this database for one of the subjects are shown in Fig. 1.5(a-d).

As sample images we will use the neutral faces, Fig. 1.5(a). As test images (i.e., images to be matched with the sample ones), we will use the happy, angry and scream faces, Fig. 1.5(b-d). For each of the test images, we will select the sample image that best matches it, as given by Eq. (2). If the retrieved image belongs to the same person (class) as the one in the testing image, we will say that our recognition was successful. Fig. 1.5(e) shows the percentage of successful identifications. We have detailed the recognition rates for each of the facial expression images to show the dependency between the recognition of identity and facial expression. We have also shown, in this figure, what would happen if the *DF* process was damaged or absent. This is represented by omitting the value of f in (1.2). The results of such damage as predicted by our model are obtained with (1.1) in Fig. 1.5(e) [50].

1.5.1.2 Computation Time As expressions increasingly diverge, the time required for the recognition of identity also increases. To calculate the time required to compute the motion field for each of the expressions, we need to determine: *i)* the number of coarse-to-fine (pyramid)levels required to compute the largest motions of the image, and *ii)* the number of iterations necessary to correctly calculate the minimum of the non-convex function at each level of the pyramid.

For each of the facial expressions in the AR database (i.e., happy, angry and scream) as well as for the neutral expression image, we have calculated the minimum number of iterations and levels of the pyramid required as follows.

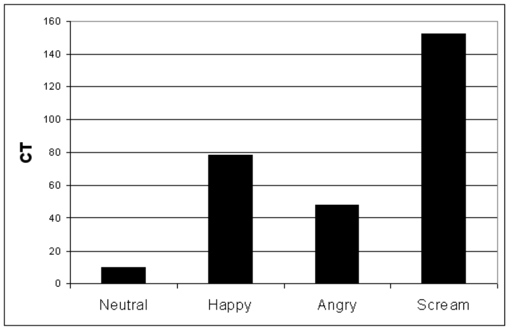


Fig. 1.6 Shown here are the mean delays (Computational Time, CT) required to compute the motion fields, f, for each facial expression group.

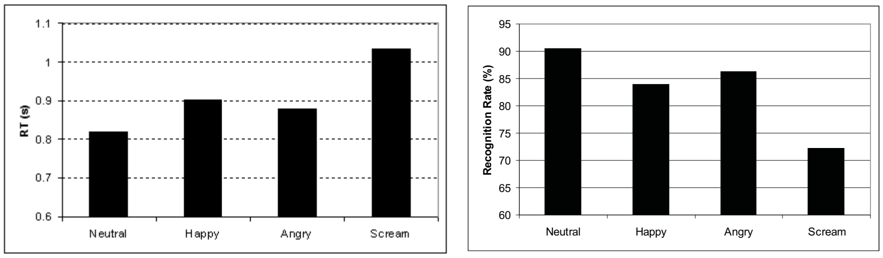
For each expression, we computed the motion fields, f, using levels of the pyramid that range from 1 to 4 . The results obtained when using $h+1$ levels of the pyramid were compared to results obtained when only using h levels. If the similarity in magnitude (as computed by S_M/m_o) and angle (S_A) between the two (h and $h+1$) was below a threshold of one pixel, we determined that h levels suffice for the computation of the motion in that image; otherwise $h+1$ levels were necessary. This chosen value is referred to as H.

To determine the number of iterations required at each level of the pyramid, we compared the results obtained when using $g+1$ and g iterations. Again, if the comparison was below a threshold, we selected g, otherwise we selected $g+1$. We will refer to this value as G. In this case, the threshold was 0.1 and g was tested for the range of values from 10 to 50.

Now, we combine the two selected values into a single measure as $CT = G \times H$, i.e., computational time = the number of iterations necessary at each level multiplied by the number of levels needed. The results (mean across samples) are: *Neutral faces*: $H = 1$, $G = 10$ and $CT = 10$, *Happy faces*: $H = 3$, $G = 26$ and $CT = 78$, *Angry faces*: $H = 2.4$, $G = 20$ and $CT = 48$, *Scream faces*: $H = 4$, $G = 33$ and $CT = 152$. These results are plotted in the graphical representation of Fig. 1.6. These results do not include the time necessary to compute (1.2), but since in our current implementation of the model this time is always constant, we can omit it for simplicity.

1.5.2 Human Performance

Subjects: Ten subjects normal or corrected-to-normal vision. *Stimuli:* Eighty (80) images of neutral, happy, angry and scream expressions of twenty (20) individuals were selected from the AR face database. To limit possible confounds, all twenty selections were males without glasses. The images were warped to a standard image (165 x 120 pixels) and displayed on a 21 inch monitor. The viewing area corresponded to approximately 15 by 11 cm. A



(a) (b)

Fig. 1.7 (a) Mean RT of the ten participants when deciding whether or not two consecutive images belong to the same individual (the prime image with a neutral expression and the target image with the expression as shown in the x-axes). (b) Mean recognition rate (in percentage).

typical viewing distance of 60 cm corresponds to 14 by 10.4 degrees of visual angle.

Design and procedure: The experiment consisted of two blocks, each with the images of ten individuals. In each block, pairs of images were shown in sequence: first a neutral image of a randomly selected individual was displayed for 800 ms (prime face), an interstimulus interval of 300 ms followed, then a neutral, happy, angry or screaming face (target face) was displayed. The identity of the prime and target face images as well as the facial expression of the target face were randomly selected. Participants were asked to decide whether or not the two images shown in sequence correspond to the same individual. Participants were instructed to respond as soon as they knew the answer. Responses and reaction times (RT) were recorded. Each subject viewed a total of 160 image pairs.

Results: Fig. 1.7(a) shows the mean RT values of all participants when deciding whether or not the prime and target face images are of the same person. As predicted, the more the target face diverged (in muscle activity) from the prime face, the greater the RT. In Fig. 1.7(b), we show the percentage in recognition rate achieved by the participants for each possible sequence pair; i.e., the prime image being a neutral expression face and the target as shown.

While the subjects' responses are predicted by our model, a numerical comparison is difficult: RTs include the matching time (which is not necessarily constant for all expressions) while the CTs correspond only to the time necessary to compute the deformation of the face (i.e. *DF* process).

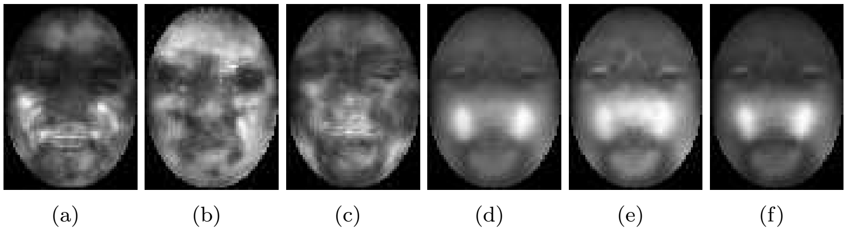


Fig. 1.8 (a) $\hat{\mathbf{S}}_{w_{happy}}$, (b) $\hat{\mathbf{S}}_{w_{angry}}$, (c) $\hat{\mathbf{S}}_{w_{scream}}$, (d) $\hat{\mathbf{S}}_b$ including all the expressions, (e) $\hat{\mathbf{S}}_b$ for expressions happy and scream, and (f) $\hat{\mathbf{S}}_b$ for expressions angry and scream.

1.6 EXPERIMENTAL RESULTS: RECOGNITION OF EXPRESSION

1.6.1 Computational Model

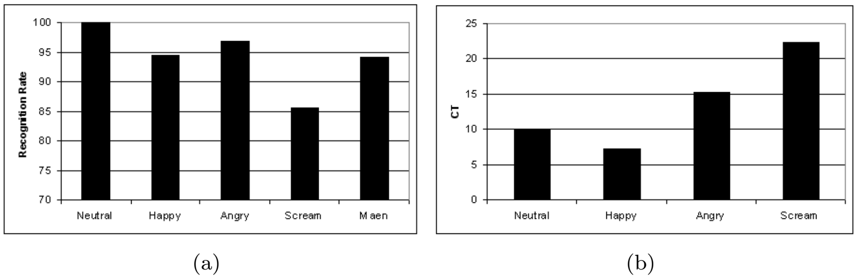
Next, we calculate the performance of the computational model in the task of expression recognition.

1.6.1.1 Performance We now show how the motion vectors can be used to recognize facial expressions. We will calculate the similarity between pairs of images by using the value of S_A described earlier in (1.12).

The first test (matching) corresponds to determining for each possible combination of two facial expressions (a total of 10 combinations) if the two images shown have the same facial expression or not. To do this, we used the neutral, happy, angry and screaming face images of 50 randomly selected individuals of the AR face database which gives us a total of $12,750$ different pairs. For each of these pairs, we compute the motion field (i.e., face deformation, *DF*) that exists between the neutral image and the facial expression selected. The two resulting motion fields are then compared by using the similarity measure S_A. This value is expected to be low for similar motion fields (i.e. similar expressions) and large for different ones.

Once the value of S_A has been obtained for each of the $12,750$ pairs of images, we search for the value of S_A that optimally divides the pairs with equal expression in one group and those with different expression within another group. We then use this threshold to classify the image pairs of a different set of 50 people. The correct classification in this second group (using the threshold obtained with the first group) was of 82.7%.

Results can be improved by means of a discriminant function that helps us to determine which areas of the face are most discriminant within classes (i.e., same facial expression) and which are most distinct between classes (i.e., different facial expressions) [51]. One way to do that is with (1.14). For instance, when comparing happy and scream faces, we can use the values of



(a) (b)

Fig. 1.9 (a) Recognition rates obtained by our model when classifying each of the face images in four different groups: neutral, happy, angry and scream. (b) Mean computational time (CT) required to calculate the class for those images with neutral, happy, angry and scream facial expressions.

$\mathbf{S}_{b(happy,scream)}$ shown in Fig. 1.8(e) and the values of $\hat{\mathbf{S}}_{w_{happy}}$ and $\hat{\mathbf{S}}_{w_{scream}}$ shown in Fig. 1.8(a,c) to determine which pixels are most discriminant, i.e., better suited for the task. We then order (rank) the pixels inversely proportional to the values of $\hat{\mathbf{S}}_w$ and proportionally to the values of $\hat{\mathbf{S}}_b$. Since most of the pixels will have an associated ranking of zero or close to zero, we can make our comparison faster by using only those pixels with a value of $\hat{\mathbf{S}}_b/\hat{\mathbf{S}}_w$ larger than a pre-determined threshold [49]. This threshold can also be learned from the training data – in which case we select that value that best classifies the training data. By following this procedure, the results improved to 91.3%.

We used the neutral, happy, angry and scream face images of 10 randomly selected individuals as samples and the neutral, happy, angry and scream face images of 90 different individuals as testing images. For each of the 360 testing images, we determine the closest sample (among the 40 stored in memory) using the value of S_A. If the facial expression in the testing image and in the closest sample were the same, we recorded a successfully classified image. Again, we use the values of $\hat{\mathbf{S}}_b$ and $\hat{\mathbf{S}}_w$ to improve the classification results and speed up computation. These results are shown in Fig. 1.9(a).

1.6.1.2 Computation Times According to our model, the delays observed when we recognize facial expressions can be due to: *i)* the time required to compute the motion field (*DF*) of the expression displayed on the (testing) image, or *ii)* the difficulty associated in classifying the facial expression of a test image in a set of pre-selected categories.

When classifying images as either happy or screaming, we expect to have longer RT for those images with a scream expression because it takes longer to compute the motion field (*DF*) of a scream face. Moreover, we would expect longer RT when classifying images as either neutral or angry than when classifying images as either happy or screaming, because the images in the first task (group) are more alike and thus a more detailed analysis will be required. While happy and screaming faces can be easily distinguish by looking at a small number of pixels (such as the eyes or the corners of the mouth), a pixel-to-pixel comparison may be necessary to decide whether an image is a neutral expression or a not-excessively-marked angry face.

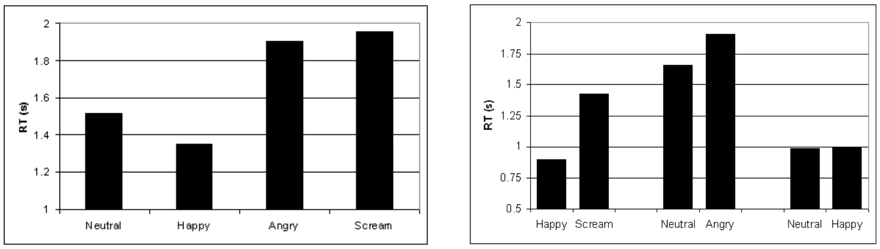
In Fig. 1.9(b) we show the Computational Times (CT) of Fig. 1.6 multiplied by the percentage (range: 0 to 1) of pixels that were necessary to use in order to obtain the best classification rate when classifying the images as either neutral expressions or the expression under consideration. The pixels were selected according to the rankings given by $\hat{\mathbf{S}}_b$.

1.6.2 Human Performance

Subjects: Ten subjects with normal or corrected-to-normal vision participated in this experiment. None of the subjects had participated in the previous experiment. *Stimuli, design and procedure:* The neutral, happy, angry and scream face images of twenty (20) males (with no glasses) of the AR face database were selected for this experiment. To prevent recognition by shape alone, images were warped to a standard image size of 165 by 120 pixels. Subjects participated in four different tests. The first required them to classify each of the images of the AR database within one of the four categories of that dataset. Subjects were told in advance of those categories and an image for each of the expressions was shown to participants before the experiment started. The other three tests only involved two types of facial expressions. In these two-class experiments, subjects were asked to classify images within these two categories only. The two-class experiments comprise the following facial expression images: a) happy and scream, b) neutral and angry, and c) neutral and happy. Reaction times (in seconds) and percentage of correct choices were recorded. Fifty images were randomly selected and displayed, one at a time, until the subject pressed a key to indicate her/his classification choice. A two second pause (with blank screen) separated each of the images shown to the participants.

Results: In Fig. 1.10(a) we show the RT means of all the participants when classifying the images within each of the four groups. These results should be compared to the CT predicted by our model and shown in Fig. 1.9(b).

As discussed above our model predicts that when classifying images into two clearly distinguishable classes, the latter will generally require longer RT because (as demonstrated in Section 3.1) longer time is required to estimate the *DF*. This was confirmed by our group of subjects, Fig. 1.10(b). We also predicted that when classifying face images within two similar classes, the RT will generally increase. This is the case for neutral and angry faces, Fig. 1.10(b). Another particular case is that of classifying face images as either neutral or happy. This task can be readily solved by looking at a small number of pixels (such as those around the corners of the lips and the eyes). Thus, in this case, similar RT are expected. This was indeed the case in our experiment, Fig. 1.10(b).



(a) (b)

Fig. 1.10 Mean RT when classifying the images in: (a) four different groups (neutral, happy, angry and scream), (b) two categories classification (happy-scream, neutral-angry, and neutral-happy). (Reaction time in seconds.)

1.7 RECOGNITION OF EXPRESSION VARIANT FACES USING WEIGHTED SUBSPACE

Zhang and Martinez [73] applied the face recognition model presented above to the subspace approach for the recognition of identity under varying expressions in the appearance-based framework. By appearance-based it is understood that the recognition system only makes use of the textural information of the face after this has been warped to a standard (prototypical) shape. Over the years, the success of appearance-based approaches, especially when applied to face recognition problems, has only increased. Appearance-based methods are attractive because the model of each class is directly defined by the selection of the sample images of that object, without the need to create precise geometrical or algebraic representations [47]. The clear disadvantage is that any image condition not included in the learning set will cause incorrect recognition results. In the pattern recognition community, it is common practice to use a minimum number of independent sample vectors of ten times the number of classes by the number of dimensions of our original feature space. Unfortunately, it is rarely the case where one has access to such a large number of training images per class in applications such as face recognition. And, even when one does have a large number of training images, these are not generally uncorrelated or independent from each other. Hence, other solutions must be defined.

The problem with subspace techniques is that some of the learned features (dimensions) represent (encode) facial expression changes. As shown above, this problem can be resolved if we learn which dimensions are most affected by expression variations and then build a weighted-distance measure that gives less importance to these. In this formulation, a fundamental question is yet to be addressed: *Would a morphing algorithm solve the problem?* That is, rather than designing a weighted measure as we did in our model, one could utilize the motion estimation to morph the test face to equal in shape that of the sample face image. This would allow a pixel to pixel comparison. Unfortunately, morphing algorithms can fail due to occlusions (e.g., teeth and closed eyes), large deformations and textural changes due to the local deformation of the face. The last of these points is key. We note that when the face changes expression, the 3D position of several local areas also change and, therefore, the reflectance angle will also change. This effect will obviously change the brightness of the image pixels (that is, the texture) in our image. The approach presented in this chapter solves this by assigning low weights to those areas with large deformations.

1.7.1 Learning linear subspace representation

Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Linear Discriminant Analysis (LDA) are three of the most popular linear subspace methods and have been largely used in face recognition applications.

PCA finds the optimal liner projection between the original space of d dimensions and a low-dimensional space of p dimensions (features) assuming the data is Gaussian [39]. To do this, PCA uses the first and central moments of the data, i.e., the sample mean μ and the sample covariance matrix $\mathbf{\Sigma}$. While PCA only computes the first and central moments of the data, ICA will use higher moments of the data to find those feature vectors that are most independent from each other [38]. In contrast, and as already described earlier in this chapter, LDA selects those basis vectors that maximize the distance between the means of each class and minimizes the distance between the samples in each class and its corresponding class mean [27].

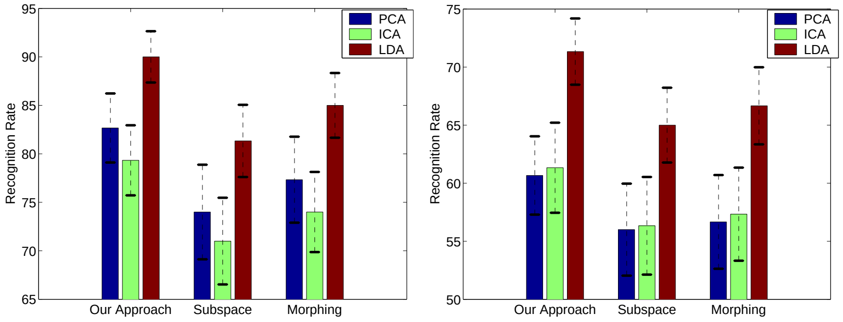
1.7.2 Weighted subspaces

Let the projection matrix given by each of the subspace methods mentioned in the preceding section be Φ_{PCA}, Φ_{ICA} and Φ_{LDA}. In this common notation, the columns in Φ_i correspond to the basis vectors of the subspace. Once these subspaces have been obtained from a training set, $\mathbf{V} = \{\mathbf{V}_1, \dots, \mathbf{V}_n\}$, n the number of training images, one can compare a new test image $\mathbf{T}$ using the following weighted-distance equation

$$\|\widehat{\mathbf{W}}_i\,(\hat{\mathbf{V}}_i - \hat{\mathbf{T}})\|, \tag{1.15}$$

where $\hat{\mathbf{V}}_i = \Phi^T\mathbf{V}_i$ which is the i^{th} image projected onto the subspace of our choice of Φ, $\Phi = \{\Phi_{PCA}, \Phi_{ICA}, \Phi_{LDA}\}$, $\hat{\mathbf{T}} = \Phi^T\mathbf{T}$, and $\widehat{\mathbf{W}}_i$ is the weighting matrix that defines the importance of each of the basis vectors in the subspace spanned by Φ. This is a direct adaptation of our model defined in (1.2) to the subspace method.

Before one can use (1.15), we need to define the value of the weighting matrix $\widehat{\mathbf{W}}$. While it may be very difficult to do that in the reduced space spanned by Φ, it is easy to calculate this in the original space and then project



(a) (b)

Fig. 1.11 Recognition rates on the leave-one-expression-out test with: (a) images from the same session, and (b) images from different sessions.

the result onto its corresponding subspace. Thus, we will compute the value of the weights in the original space, $\mathbf{W}$, using of the model given in section 1.4.2 yielding $\mathbf{F}_i = DF(\mathbf{V}_i, \mathbf{T})$. The weights are given by

$$\mathbf{W}_i = \mathbf{F}_{max} - \|\mathbf{F}_i\|, \tag{1.16}$$

where $\mathbf{F}_{max} = max_i \|\mathbf{F}_i\|$.

We can now project the weights onto the corresponding subspace as

$$\hat{\mathbf{W}}_i = \Phi^T \mathbf{W}_i. \tag{1.17}$$

To classify a test image, we assign the class label of the closest sample. That is given by,

$$s = argmin_i \|\hat{\mathbf{W}}_i \left(\hat{\mathbf{V}}_i - \hat{\mathbf{T}} \right) \|, \tag{1.18}$$

and selecting the class label, c^*, of $\mathbf{V}_s$.

1.7.3 Experimental results

Once more, we randomly selected 100 subjects from the AR face database. From each individual we used the eight images with neutral, happy, angry and scream expressions taken during two sessions – each session separated by two weeks time [46]. All algorithms were tested using the leave-one-expression-out procedure. For example, when the happy face was used for testing, the neutral, angry and scream faces were used for training.

In our first experiment, only those images taken during the first session were used. The results obtained using the proposed weighted subspace approaches as well as those of PCA, ICA and LDA are shown in Fig. 1.7.3(a). In this figure we also show the results obtained by first morphing all faces to a neutral-expression face image and then building a PCA, ICA and LDA subspace. The bars in this figure show the average recognition rate for each of the methods. The standard deviation for the leave-one-expression-out test is shown by the small variance line at the top of the bars.

The second test was similar to the first one, except that, this time, we used the images of the first session for training and those of the second session for testing. The results are summarized in Fig. 1.7.3(b). It is worth mentioning that the weighted-LDA approach works best for the scream face with a recognition rate of $\sim 84\%$. Other methods could not do better than 70% for this expression. In the figures shown above this is made clear by the small variance associated to our method as compared to the others.

1.7.4 Recognition from Video Sequences

Compared to the large number of algorithms developed to do recognition from still images, the literature on video-based methods is relatively small. One reason for this imbalance was due to the low accessibility of high-quality video cameras which, until recently, were expensive and of limited quality. The second reason is algorithmical. While it is generally difficult to successfully do feature extraction from still images, this process has proven even more challenging over dynamic sequences [51]. This second point raises an important question: *Would the methods defined to recognize faces from a single test image perform better if they could work with multiple images or video sequences?* Note that if the answer to this question were positive, there would be less need for the design of feature extraction algorithms that can do a more direct analysis of dynamic sequences. Understanding the limitations of current algorithms when applied to video will help researchers design algorithms that can specifically solve these problems [74].

To answer our question though, we need to be able to use our computational model, originally defined to work with stills, to handle multiple images. Zhang and Martinez [74] present one such approach. In their algorithm the method of [48] is reformulated within the framework presented in this chapter. This results in a robust algorithm that can accurately recognize faces even under large expressions, pose and illumination changes and even under partial occlusions. Experimental results using a database of video sequences corresponding to fifty people yielded a classification accuracy of $\sim 95\%$.

1.8 SUMMARY

In the model presented in this chapter, and depicted in Fig. 1.3, motion (dynamic) cues are processed independently from static cues. This is consistent with neurophysiological evidence that supports dorsal stream processing of dynamic cues and ventral stream processing of static cues. Although dynamic and static cues are processed separately in our model, they are combined to accomplish the tasks of recognition of identity and facial expression at the end of the hierarchy. This is also consistent with experimental data that show disruption in recognition when one of the two cues (dynamic or static) is altered [49].

The nature of this model provides a framework in which to reconcile apparently contradictory psychophysical data. The process of motion estimation (whose task is to calculate the deformation between the faces we want to match), *DF*, within a hierarchical model of face processing is key to explaining why in some experiments – slower recognition times are obtained when attempting to identify faces with distinct facial expression, e.g., smiling versus neutral faces. At the same time, the model does not require a direct interaction between the processes of face identification and facial expression recognition. This is important, because it is consistent with the observation that some agnosic patients are impaired only with regard to one of the two tasks (either identification of people or facial expression recognition).

This model suggests that motion is useful for successful matching of face images bearing distinct facial expressions. Following [49, 50], we further hypothesized that the computed motion fields could be used to select the most invariant textural (appearance) features between the images we want to match. It is observed that the results that generated by the proposed model is consistent with psychophysical and neurophysiological data. Additionally, recognition of identity is reduced by discarding the outcome of the *DF* module from the similarity function, i.e., going from (1.2) to (1.1). These motion features could also be used to construct a motion-based feature-space for the recognition of identity and expression. Motion may be used as an alternative, independent means for identifying people and expressions. In computer vision, reasonable results have been obtained by constructing feature-spaces based solely on motion cues. These results could ultimately be used to reinforce the recognition task, or help to make a decision where other processes are not adequate.

We have demonstrated the use of our model to classify faces within a set of facial expression categories. We have also experimentally shown that the *DF* carries the necessary information to successfully achieve this task. By combining the *DF* and a linear classifier, we were able to predict the classification RT of each of the facial expressions of the AR database.

Extensions to the classical subspace approach [73] and to the recognition from video sequences [74] show the generality of the model presented in this chapter. Moreover, this chapter has illustrated how one can successfully employed the model defined herein to make predictions on how the human visual system works – predications later confirmed in a set of psychophysical experiments.

REFERENCES

1. R.D. Adams and M. Victor, *Principles of Neurology.* New York: McGraw-Hill (1981).

2. T. Allison, A. Puce and G. McCarthy, Social perception from visual cues: role of the STS region. *Trends in Cognitive Science*, **4**:267-278 (2000).

3. M.S. Bartlett, J.C. Hager, P.Ekman, and T.J. Sejnowski, Measuring spatial expressions by computer image analysis. *Psychophysiology*, **36**:253-263 (1999).

4. J. Baudouin, D. Gilibert, S. Sansone, and G. Tiberghien, When the smile is a cue to familiarity. *Memory*, **8**:285-292 (2000).

5. M. Behrmann and G. Avidan, Congenital prosopagnosia: face-blind from birth. *Trends in Cognitive Sciences*, **9**, 180-187 (2005).

6. D. Beymer and T. Poggio, Face recognition from one example view. *Science*, **272**(5250) (1996).

7. D. Bimler and J. Kirkland, Categorical perception of facial expressions of emotion: Evidence from multidimensional scaling. *Cognition and Emotion*, **15**, 633-658 (2001).

8. M.J. Black and P.Anandan, The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, **63**:75-104 (1996).

9. A. Blake and A. Zisserman, *Visual reconstruction.* Cambridge: The MIT Press (1987).

10. M.H. Bornstein, Perceptual categories in vision and audition. In *Categorical Perception: the groundwork of cognition*, S. Harnad (Ed.), Cambridge University Press (1987).

11. V. Bruce and A. Young, Understanding face recognition. *British Journal of Psychology*, **77**:305-327 (1986).

12. R. Bruyer, C. Laterre, X. Seron, P. Feyereisen, E. Strypstein, E. Pierrard and D. Rectem D. (1983). A case of prosopagnosia with some preserved covert remembrance of familiar faces. *Brain and Cognition* **2**:257-284 (1983).

13. J.T. Cacioppo, W.L. Gardner and G.G. Berntson, The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, **76**, 839-855 (1999).

14. A.J. Calder, A.W. Young, P.J. Benson and D.I. Perrett, Categorical perception of morphed facial expressions. *Visual Cognition*, **3**:81-117 (1996).

15. A.J. Calder and A.W. Young, Understanding the recognition of facial identity and facial expression. *Neuroscience*, **6**, 641-651(2005).

16. J.F. Cohn and K.L. Schmidt, The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, **2**, 1-12 (2004).

17. C. Darwin, *The Expression of the Emotions in Man and Animals*. London:John Murray, 1872.(Re-printed by TheUniversity of Chicago Press 1965.)

18. M. de Haan, K. Humphreys and M.H. Johnson, Developing a brain specialized for face perception: A converging methods approach. *Developmental Psychobiology*, **40**, 200-212 (2002).

19. P. Ekman, Basic Emotions. In T. Dalgleish and M. Power (Eds) *Handbook of Cognition and Emotion* New York: Wiley & Sons (1999).

20. P. Ekman, R.J. Davidson and W.V. Friesen, The Duchenne Smile: emotional expression and brain physiology II. *Journal of Personality and Social Psychology* **58**, 342-353 (1990).

21. P. Ekman and W.V. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement* San Diego: Consulting Psychology Press (1978).

22. A.W. Ellis, Cognitive mechanisms of face processing. *Philosophical Transactions of the Royal Society of London B* **335**, 113-119 (1992).

23. H.D. Ellis and M.B. Lewis, Capgras delusion: a window on face recognition. *Trends in Cognitive Sciences*, **5**, 149-156(2001).

24. N. Endo, M. Endo, T. Kirita and K. Maruyama, The Effects of Expression on face Recognition. *Tohoku Psychologia Folia*, **52**:37-44 (1992).

25. N.L. Etcoff, Selective attention to facial identity and facial emotion. *Neuropsychologia*, **22**:281-295 (1984).

26. M.J. Farah, *Visual Agnosia: Disorders of object recognition and what they tell us about normal vision.* Cambridge: MIT Press (1990).

27. R.A. Fisher, The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**:179-188A (1936).

28. V. Gallese, C. Keysers and G. Rizzolatti, A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, **8**(9), 396-403 (2004).

29. I. Gauthier, M.J. Tarr, A.W. Anderson, P. Skudlarskiand J.C. Gore, Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, **2**, 568-573 (1999).

30. E.C. Hansch and F.J. Pirozzolo, Task Relevant Effects on the Assessment of Cerebral Specialization for Facial Emotions. *Brain and Language* **10**:51-59 (1980).

31. S. Harnad, *Categorical Perception* New York: Cambridge University Press (1987).

32. J.V. Haxby, E.A. Hoffman and M.I. Gobbini, The distributed human neural system for face perception. *Trends in Cognitive Science*, **4**:223-233 (2000).

33. D.C. Hay, A.W. Young and A.W. Ellis, Routes through the face recognition system. *Quarterly Journal of Experimental Psychology A*, **43**:761-791(1991).

34. H. Hill and A. Johnston, Categorizing sex and identity from the biological motion of faces. *Current Biology*, 11:880-885 (2001).

35. E.A. Hoffman and J.V. Haxby, Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, **3**, 80-84 (2000).

36. B.K.P. Horn and B.G. Schunck, Determining optical flow. *Artificial Intelligence*, **17**:185-203 (1981).

37. G.W. Humphreys, N. Donnelly and M.J. Riddoch, Expression is computed separately from facial identity, and it is computed separately for moving and static faces – Neuropsychological evidence. *Neuropsychologia*, **31**:173-181 (1993).

38. A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis. Wiley-Interscience, 2001.

39. I.T. Jolliffe, Principal Component Analysis. Springer-Verlag, 2002.

40. M. Kamachi, V. Bruce, S. Mukaida, J. Gyoba, S. Yoshikawa and S. Akamatsu, Dynamic properties influence the perception of facial expressions. *Perception*, **30**:875-887 (2001).

41. N. Kanwisher, J. McDermott, and M.M Chun, The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, **17**, 4302-4311 (1997).

42. Y. Kim, A.M. Martinez, and A.C. Kak, Robust Motion Estimation under Varying Illumination. *Image and Vision Computing*, **23**, 365-375 (2005).

43. C.L. Krumhansl, Music psychology: Tonal structures in perception and memory. *Annual Review of Psychology*, **42**, 277-303 (1991).

44. J. Kurucz and G. Feldmar, Prosopo-affective agnosia as a symptom of cerebral organic-disease. *Journal of American Geriatric Society*, **27**:225-230 (1979).

45. K. Lander, F. Christie and V. Bruce, The role of movement in the recognition of famous faces. *Memory Cognition*, **27**:974-985 (1999)

46. A.M. Martinez and R. Benavente, The AR-Face Database. *CVC Tech. Report #24*(1998).

47. A.M. Martinez and A.C. Kak, PCA versus LDA. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **23**(2):228-233 (2001).

48. A.M. Martinez, Recognizing Imprecisely Localized, Partially Occluded and Expression Variant Faces from a Single Sample per Class. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **24**(6):748-763 (2002).

49. A.M. Martinez, Matching expression variant faces. *Vision Research*, **43**, 1047-1060 (2003).

50. A.M. Martinez, Recognizing Expression Variant Faces from a Single Sample Image per Class. *In Proceedings of IEEE Computer Vision and Pattern Recognition*, **1**, 353-358 (2003).

51. A.M. Martinez, and M. Zhu, Where Are Linear Feature Extraction Methods Applicable? *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **27**(12):1934-1944 (2005).

52. D. Maurer, R. Le Grandand C.J. Mondloch, The many faces of configural processing. *Trends in Cognitive Sciences*, **6**, 255-260 (2002).

53. S. Negahdaripour, "Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, 20(9):961-979, 1996.

54. C.A. Nelson, The development and neural bases of face recognition. *Infant and Child Development*, **10**, 3-18 (2001).

55. M. Pantic and L.J.M. Rothkrantz, Automatic analysis of facial expressions: the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **22**(12):1424-1445 (2000).

56. M.T. Posamentier and H. Abdi, Processing faces and facial expressions. *Neuropsychology Review*, **13**, 113-143 (2003).

57. D.A. Roark, S.E. Barrett, M.J. Spence, H. Abdi and A.J. O'Toole, Psychological and neural perspectives on the role of motion in face recognition. *Behavioral and Cognitive Neuroscience Reviews*, **2**, 15-46 (2003).

58. E.T. Rolls, A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion*, **4**, 161-190 (1990).

59. J.A. Russell, A Circumplex model of affect. *J. Personality and Social Psych.*, **39**:1161-1178 (1980).

60. J.A. Russell and M. Bullock, Multidimensional scaling of emotional facial expressions: similarity from preschoolers to adults. *Journal of Personality and Social Psychology*, **48**, 1290-1298(1985).

61. J.A. Russell, Core affect and the psychological construction of emotion. *Psychological Review*, **110**, 145-172 (2003).

62. K.L. Schmidt and J.F. Cohn, Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology*, **S33**, 3-24 (2001).

63. S.R. Schweinberger and G.R. Soukup, Asymmetric relationship among perception of facial identity, emotion, and facial speech. *J. Exp. Psychology: Human Perception and Performance*, **24**(6):1748-1765 (1998).

64. M.J. Tarr and I. Gauthier, FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*, **3**, 764-769 (2000).

65. L.G. Ungerleider, S.M. Courtney, and J.V. Haxby, A neural system for human visual working memory. *Proceedings of the National Academy of Science of USA*, **95**, 883-890 (1998).

66. T. Valentine, Face-space models of face recognition. In M.J. Wenger and J.T. Townsend (Eds.) *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges.* London: Lawrence Erlbaum Associates Inc.

67. G. Wallis and H.H. Bulthoff, Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. USA*, **98**:4800-4804 (2001).

68. R.S. Woodworth and H. Schlosberg, *Experimental Psychology.* New York: Holt, Rinehart, & Winston (1954).

69. Y. Yacoob and L. Davis, Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **18**:636-642 (1996).

70. A.W. Young, D.J. Hellawell, C. Van De Wal and M. Johnson, Facial expression processing after amygdalotomy. *Neuropsychologia*, **34**:31-39 (1996).

71. A.W. Young, D. Rowland, A.J. Calder, N.L. Etcoff, A. Seth and D.I. Perrett, Facial expression megamix: Test of dimensional and category accounts of emotion recognition. *Cognition*, **63**:271-313 (1997).

72. G. Yovel and N. Kanwisher, Face perception: Domain specific, not process specific. *Neuron*, **44**, 889-898 (2004).

73. Y. Zhang and A.M. Martinez, Recognition of Expression Variant Faces Using Weighted Subspaces. *Proceedings of International Conference on Pattern Recognition (ICPR)*, **3**, 149 - 152 (2004).

74. Y. Zhang and A.M. Martinez, A Weighted Probabilistic Approach to Face Recognition from Multiple Images and Video Sequences. *Image and Vision Computing*, **24**, 626-638 (2006).

75. M. Zhu and A.M. Martinez, Subclass Discriminant Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **28**(8):1274-1286 (2006).